Homework # 8
Due 3/13

1. Let $f(x,y) = 100(y-x^2)^2 + (1-x)^2$. This is the famed banana function (see wikipedia). The minimum of f is $(1,1)$ (why?). Here we know the minimum, so we don't need to numerically optimize, but the banana function serves as a simple setting to compare gradient descent to Newton's method.

   (a) Starting at the point $(4,4)$ use a gradient descent with a fixed learning rate to locate the minimum. How many iterations before you find the minimum? You can alter the learning rate to get the quickest convergence you can. You can also vary the learning rate over the iterations.

   (b) Now repeat but use Newton's method. How many iterations now?

2. Here we'll use a logistic regression to build a classifier for the MNIST dataset. To keep things manageable, let's build a classifier that determines whether an image is a 3 or not a 3. (If we have a chance in subsequent homeworks, we'll build a classifier for all the digits.) For each MNIST image x_i, let $y_i = 1$ if the ith image is a 3, otherwise $y_i = 0$. The $x_i \in \mathbb{R}^{784}$, but add a leading 1 to each image to make $x_i \in \mathbb{R}^{785}$. (This will allow the linear function $\alpha \cdot x$ below to have a non-zero intercept term.) Recall the logistic regression model,

$$P(y = 1|x, \alpha) = \frac{1}{1 + \exp(-\alpha \cdot x)}, \tag{1}$$

   where $x \in \mathbb{R}^{785}$ is an MNIST image. Define the likelihood function,

$$L(\alpha) = \prod_{i=1}^{n} P(y_i \mid x_i, \alpha), \tag{2}$$

   where n is the number of images.

   (a) Show that $\ell(\alpha) = \log L(\alpha)$ satisfies,

$$\ell(\alpha) = \sum_{i=1}^{N} (1 - y_i)(-\alpha \cdot x_i) - \log(1 + \exp(-\alpha \cdot x_i))) \tag{3}$$

(b) Derive a formula for the gradient and Hessian of $\ell(\alpha)$. Show that the Hessian is semi-negative definite (i.e. all eigenvalues are non-positive).

(c) Split the MNIST dataset into a training dataset and a test dataset. Using the training dataset, apply steepest ascent to find the α that maximizes $\ell(\alpha)$. Then use Newton's method on the training dataset to find the optimal α. You will find that the Hessian may not be invertible. In this case consider perturbing the Hessian to $-\rho I + H$ where $\rho > 0$. Explain why this will make the Hessian invertible. Use this perturbed matrix instead of the Hessian in Newton's method.

(d) Given the optimal α that you computed above, construct a classifier that determines if an image is a 3 or not. Apply your classifier to the test dataset and determine your accuracy.